DOCUMENT RESUME

ED 241 797 CE 038 660

AUTHOR	Burnside, Billy L.
TITLE	Subjective Appraisal as a Feedback Tool. Technical Report 604.
INSTITUTION	Army Research Inst. for the Behavioral and Social Sciences, Alexandria, Va.
PUB DATE	May 82
NOTE	44p.; Developed at ARI Field Unit, Fort Knox, KY.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Evaluation; Evaluation Methods; Feedback; Informal Assessment; *Interviews; *Job Performance; *Military Personnel; Military Training; Peer Evaluation; *Personnel Evaluation; Self Evaluation (Individuals); *Surveys
IDENTIFIERS	*Accuracy; *Subjective Evaluation

ABSTRACT

This report examines the accuracy of subjective appraisals of several aspects of task performance, including proficiency, difficulty, frequency, and criticality An introduction discusses current Army use of subjective appraisal, feedback methods, and problems with subjective appraisal. Data pertaining to the accuracy of various types of appraisal are summarized in the next section. The types of appraisal included are proficiency appraisals, task criticality appraisals, task difficulty appraisals, task frequency estimates, and appraisal of training materials. At the end of the section, research from the cognitive psychology literature relating to human ability to make accurate subjective appraisals is discussed. The third section summarizes data relating to the relative accuracy of appraisals obtained from alternative sources, namely supervisor, self-, and peer appraisals. In the fourth section, discussion of the issue of how subjective appraisals are formulated centers around survey and interview techniques. Methods for increasing the accuracy of subjective appraisals are made, focusing on phrasing of questions, raters' experience, and other characteristics of raters. The paper concludes with suggestions for optimizing combined use of the survey and interview approaches. (YLB)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED241797

Technical Report 604

SUBJECTIVE APPRAISAL AS A FEEDBACK TOOL

Billy L. Burnside

ARI FIELD UNIT AT FORT KNOX, KENTUCKY

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

U. S. Army

Research Institute for the Behavioral and Social Sciences

May 1982

Approved for public release; distribution unlimited.

CE038660

U. S. ARMY RESEARCH INSTITUTE
FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency under the Jurisdiction of the

Deputy Chief of Staff for Personnel

JOSEPH ZEIDNER
Technical Director

L. NEALE COSBY
Colonel, IN
Commander

NOTICES

DISTRIBUTION: Primary distribution of this report has been made by ARI. Please address correspondence concerning distribution of reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, ATTN: PERI-TST, 5001 Eisenhower Avenue, Alexandria, Virginia 22333.

FINAL DISPOSITION: This report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Technical Report 604	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) SUBJECTIVE APPRAISAL AS A FEEDBACK TOOL		5. TYPE OF REPORT & PERIOD COVERED Interim Report October 1981 - May 1982
		6. PERFORMING ORG. REPORT NUMBER ---
7. AUTHOR(s) Billy L. Burnside (ARI)		8. CONTRACT OR GRANT NUMBER(s) ---
9. PERFORMING ORGANIZATION NAME AND ADDRESS US Army Research Institute for the Behavioral and Social Sciences, 5001 Eisenhower Avenue Alexandria, VA 22333		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 2Q263743A794
11. CONTROLLING OFFICE NAME AND ADDRESS US Army Research Institute for the Behavioral and Social Sciences, 5001 Eisenhower Avenue Alexandria, VA 22333		12. REPORT DATE May 1982 13. NUMBER OF PAGES 43
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) ---		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE ---

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

---

18. SUPPLEMENTARY NOTES

---

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Training Feedback
Training Effectiveness
Training Management
Training Development
Training Evaluation
Human Performance
Instructional Design

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This report examines the accuracy of subjective appraisals of several aspects of task performance, including proficiency, difficulty, frequency, and criticality. The relative accuracy of subjective appraisals collected from various sources by various methods is discussed, and suggestions are developed for ways to increase the accuracy of these appraisals. The use of subjective data in an integrated feedback system is addressed, and suggestions for further research are offered. Findings should be of interest to training developers and evaluators.

Technical Report 604

SUBJECTIVE APPRAISAL AS A FEEDBACK TOOL

Billy L. Burnside

Submitted by:
Donald F. Haggard, Chief
ARI FIELD UNIT AT FORT KNOX, KENTUCKY

Approved by:
Harold F. O'Neil, Jr., Director
TRAINING RESEARCH LABORATORY

U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES
5001 Eisenhower Avenue, Alexandria, Virginia 22333

Office, Deputy Chief of Staff for Personnel
Department of the Army

May 1982

Army Project Number
2Q263743794

Education and Training

Approved for public release; distribution unlimited.

ARI Research Reports and Technical Reports are intended for sponsors of R&D tasks and for other research and military agencies. Any findings ready for implementation at the time of publication are presented in the last part of the Brief. Upon completion of a major phase of the task, formal recommendations for official action normally are conveyed to appropriate military agencies by briefing or Disposition Form.

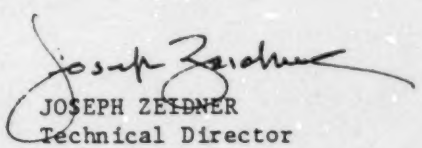
FOREWORD

The Fort Knox Field Unit has long been involved in the application of experimental psychology to increasing the quality of the products of Army Centers/Schools. These products include trained soldiers and training materials. The training evaluation and feedback team of this unit performs research and development on increasing the quality of these products by improving the information flow between training developers and users in the field.

In ARI Research Report 1323 (Burnside, 1981), it was determined that the principal methods currently used to provide feedback from field personnel to training developers involve the collection of subjective data. Such data involve individuals' judgments or estimations, which may or may not be objectively verified in particular instances. This approach is a cost-effective one, but the accuracy of the data involved is a matter for concern. This issue of the accuracy of subjective data must be resolved before an integrated feedback system can be designed. Decisions must be made as to when subjective data can be relied upon and when more objective but costly methods must be applied.

This report provides background for the integration of subjective and objective feedback methods by examining the accuracy of subjective data in a variety of settings. Findings indicate that such data are frequently not accurate and should be used cautiously. Included in the report are suggestions for ways to increase the accuracy of subjective data, and these have implications for TRADOC and other Army personnel concerned with the evaluation of training and the flow of information between training developers and users.

A wide range of data is summarized in this report. Significant assistance in locating many of these data was provided by the peer reviewers, Dr. Jack Hiller of the Presidio of Monterey Field Unit and Dr. Joel Schendel of the Fort Benning Field Unit. They also provided many useful comments which have been incorporated into the report. Acknowledgement is also extended to Dr. Stephen Goldberg of the Fort Knox Field Unit for the provision of unpublished data used in this report.



JOSEPH ZEIDNER
Technical Director

SUBJECTIVE APPRAISAL AS A FEEDBACK TOOL

EXECUTIVE SUMMARY

Requirement:

Feedback from field units to US Army Training and Doctrine Command (TRADOC) Centers/Schools currently consists largely of subjective data, or information which may be influenced by individuals' opinions or inferences. In this report the accuracy of such data is examined in order to determine their utility as a feedback tool.

Procedure:

Relevant previously published and unpublished data are reviewed from a variety of sources, including military research, educational research, and cognitive psychology. These data are organized to address the accuracy of subjective appraisals of individuals' proficiencies on specific tasks, as well as the task performance frequency, difficulty, and criticality. Other issues addressed are the relative accuracy of various sources of subjective appraisals (self, supervisors, and peers) and the relative accuracy of various appraisal methods (survey and interview techniques).

Findings:

Subjective appraisals of various aspects of task performance have been found to be accurate in some instances. But, in general, accuracy of subjective appraisals has not been reported consistently enough to support their widespread use as feedback without further accuracy checks. The relative accuracy of various subjective appraisal sources and methods has also not been fully determined. Various proposals for further research and for ways in which the accuracy of subjective appraisals may be increased are included in the report.

Utilization of Findings:

This report will be useful to training developers and evaluators to assist them in obtaining meaningful feedback on various aspects of task performance from the field. It will also be useful in guiding development of an integrated feedback system and in guiding research on design of cost-effective and accurate feedback tools.

SUBJECTIVE APPRAISAL AS A FEEDBACK TOOL

CONTENTS

	Page
INTRODUCTION	1
Current Army Use of Subjective Appraisal	2
Feedback Methods	3
Problems with Subjective Appraisal	3
Report Organization	4
TYPES OF APPRAISALS	4
Proficiency Appraisals	4
Task Criticality Appraisals	11
Task Difficulty Appraisals	12
Task Frequency Estimates	13
Appraisal of Training Materials	14
Tentative Conclusions	15
Cognitive Psychology	15
TYPES OF APPRAISERS	18
Relative Accuracy of Self-Appraisals	19
Peer Appraisals	22
Tentative Conclusions	22
TYPES OF APPRAISAL METHODS	23
Surveys and Interviews	23
Phrasing of Questions	24
Raters' Experiences	25
Other Characteristics of Raters	26
CONCLUSIONS/RECOMMENDATIONS	27
REFERENCE NOTES	30
REFERENCES	31

INTRODUCTION

The purpose of this paper is to determine the accuracy and utility of a particular evaluation method, subjective appraisal. Appraisal here refers to the evaluation of the performance of individual soldiers and military units on specific tasks in a field setting. This is distinguished from assessment, which involves a general evaluation of individuals' personal characteristics, knowledge, and abilities, such as the evaluation of leadership abilities in an assessment center (Levine, 1980). The present paper is primarily concerned with appraisal of task-specific job performance, and not with more general assessment issues.

The terms "subjective" and "objective" will be used frequently throughout this paper, and they have numerous connotations. It is thus necessary to define their meanings carefully in the present context. Dictionary definitions of "subjective" include "illusory" and "existing only within the experiencer's mind and incapable of external verification." Such negative connotations are not intended here. Rather, subjective appraisal is defined as that which is based upon individuals' judgments or estimations, and which can be but is not always externally verified. Subjective appraisals are usually obtained through the use of surveys or interviews, in terms of some sort of rating scale. In contrast, dictionary definitions of "objective" include "having to do with material objects, actual existence, or observable phenomena" and "uninfluenced by emotion or personal prejudice." Objective appraisal thus involves the actual observation of performance and collection of performance data; i.e., verification external to individuals' opinions or estimations. For example, one could simply ask a soldier whether he or she can perform a specific task; this is what is meant by subjective appraisal here. Or one could administer a hands-on test, observe the soldier's performance, and compare it against a validated standard; this is what is meant by objective appraisal. The distinction is analogous to that frequently made between "soft" and "hard" data, with "soft" data consisting largely of individuals' opinions and intuitive judgments and "hard" data consisting of performance results in a controlled situation. Objective appraisal (or "hard" data) provides in some sense the truest evaluation, since it is observable and externally verified. But subjective appraisal (or "soft" data) is the more efficient and cost-effective method. In some real-world situations, objective appraisal may be so costly and time-consuming as to be practically impossible. A key question then becomes that of whether data gathered during subjective appraisal are sufficiently accurate to warrant their use in particular situations. This is a primary issue in the present paper.

In actuality, the distinction between subjective and objective appraisal is not as clear-cut as might have been implied above. Appraisal is perhaps best described as a dimension with subjectivity at one end and objectivity at the other. The difference between subjective and objective appraisal methods is thus one of degree, with real-world methods representing various mixes. Surveys can be made more objective by asking well-specified factual questions and by using behaviorally anchored rating scales (Cascio, 1978). Performance observation can be made less objective by using written knowledge tests or simulated performance in lieu of actual "hands-on" performance, or by using

observational criteria or standards which require judgments or inferences to be made. One could enter into protracted philosophical arguments about the distinction between subjective and objective appraisal; all subjective opinion is based upon experience to some extent, and all objective performance observation and testing involves judgment to some extent. Such arguments will be avoided here in the interest of practicality. For practical purposes, the key question is not how the methods differ in a theoretical sense or whether one method is better than the other in an absolute sense, but rather what the appropriate mix of methods is for a given situation.

Current Army Use of Subjective Appraisal

The use of subjective appraisal and assessment methods is ubiquitous in the Army. The career performance of individuals is periodically assessed with efficiency reports which utilize subjective rating scales and narrative comments. The readiness of units is periodically assessed using a Unit Status Report (AR 220-1) which requires subjective estimates on the part of the unit commander (Heymont, 1977). The collective performance of units on specific exercises, such as Table IX for tank platoons and Army Training and Evaluation Program (ARTEP) missions, is largely appraised subjectively because the complexity of the performance would make objective appraisal highly resource-intensive. Task analyses and front-end analyses for new training programs are frequently based upon subjective appraisals. For example, subjective estimates of the criticality and performance frequency of specific tasks may be obtained by administering Comprehensive Occupational Data Analysis Program (CODAP) surveys to field personnel. Problem-solving techniques used in the Army, such as the estimate of the situation (FM 101-5), also frequently require the subjective appraisal of specific situations and courses of action. The use of subjective appraisal is so widespread in the Army that it has in some respects been canonized, is commonly referred to as "military judgment," and is sometimes espoused by senior Army personnel as the only approach for analyzing complex situations (West, Note 1).

The scope of this paper does not allow a review of subjective judgment in the Army in all its manifestations. Rather, the use of subjective appraisal will be examined in a specific context or situation, the feedback of information from field units to Centers/Schools. The products of Training and Doctrine Command (TRADOC) Centers/Schools can be grouped into two categories: graduates and training doctrine, guidance, or materials. In order to appraise the quality and utility of these products, elements of the Centers/Schools need meaningful feedback from users in the field. This constitutes the evaluation phase of the Instructional Systems Development (ISD) process described in TRADOC Pam 350-30 and further delineated in draft TRADOC Regulation 350-7. Elements collecting feedback from users may include Directorates of Evaluation (DOE's), task analysts, training developers, and special offices (e.g., the Office of Armor Force Management and Standardization (OAFMS) at Fort Knox, KY). A preliminary review indicated that the primary methods which such elements currently use to gather feedback frequently include the use of subjective appraisals (Burnside, 1981).

Feedback Methods

There are six principle methods which Centers/Schools may use to obtain feedback from field units: receipt of informal comments, administration of surveys/questionnaires, conduct of interviews, analysis of existing unit performance records, observation of field performance, and operational field performance testing. The first three of these methods, which definitely involve subjective appraisal, are the most frequently used according to battalion commanders and staffs (Burnside, 1981). The last two methods are more objective in nature, but are not commonly used because of their costs. The sixth method, analysis of existing records, may best be described as a mix of subjective and objective appraisal, but it was found to be of limited utility because of the limited availability, standardization, and specificity of many records. Burnside (1981) reviewed the general parameters and usage of available feedback methods. The present paper provides further analysis of the accuracy of the most popular of these methods; i.e., those involving subjective appraisal.

Problems with Subjective Appraisal

What are the general problems which may arise from the use of subjective appraisal? Reviewers of the subjective judgment literature (e.g., Cascio (1978), Holzbach (1978), and Thornton (1980)) have consistently described several types of psychometric errors or problems which commonly occur. Prominent among these are leniency errors, central tendency errors, halo effects, and lack of interrater reliability. Leniency errors occur when raters avoid using the low extremes of a rating scale, leading to a restricted range or reduced variance of ratings. This tendency may represent a systematic bias on the part of raters to avoid giving ratings which can be interpreted negatively. The occurrence of leniency errors among Army raters is exemplified by past distributions of officer efficiency ratings, in which only the top few points of a 100-point scale have been used. Similar to leniency errors are central tendency errors, which represent a tendency of raters to avoid using both the high and low extremes of rating scales. If there is no systematic bias against negative ratings, there may still be a bias against extreme ratings and a tendency for responses or judgments to cluster around the middle of the scale. Thus, everything is rated about average, and the variance of ratings is again reduced.

The halo effect occurs when a rater fails to distinguish among the different dimensions of a situation and applies a global or overall judgment based on one salient dimension. The ratings of different aspects of a situation then tend to agree or correlate highly, whether this is appropriate or not. For example, if a supervisor is asked to rate the performance of a soldier on specific tasks, he or she may make the global judgment that the soldier is a good worker and rate him or her high on all tasks, even though performance of some of them may never have been observed. Such a rating tendency detracts from the ability to discriminate between different aspects of performance.

The lack of interrater reliability simply means that different raters do not agree in their judgments. Without reliability, ratings are practically useless; reliability sets the limit on the degree of validity which can be

obtained (Mitchell, 1979). For example, if a group of subject matter experts do not agree on ratings of task criticality, then "truly" critical tasks cannot be identified. Of course, agreement among raters does not guarantee accuracy of ratings (Frick and Semmel, 1978). Raters can all agree and all be wrong. So interrater reliability is a necessary but not sufficient prerequisite to obtaining valid ratings.

One effect of the problems briefly described above is to reduce the amount of correlation or agreement between subjective ratings and more objective criteria. For example, a tendency which reduces the variance of ratings generally reduces the degree to which they correlate with other measures. These and other problems with the use of subjective ratings in feedback will be further discussed in the context of specific sample data below. Approaches for eliminating or reducing rater bias will be addressed in the final section of the paper.

Report Organization

There are numerous dimensions or sets of issues which could be used to organize discussion of the area of subjective feedback. The organization used in this report will center around the issues of what is being appraised, who is doing the appraising, and how the appraisal is being done. The type of appraisal of greatest interest here involves estimates of soldiers' proficiencies on specific tasks. But other types of appraisals are of interest to TRADOC Centers/Schools, at least during front-end analysis, and these include judgments of the criticality, difficulty, and performance frequency of specific tasks. Data pertaining to the accuracy of all these types of appraisals are summarized in the next section. With regard to the issue of who performs subjective appraisals, the most common approaches in the feedback arena are self-appraisal and appraisal by supervisors. Another approach which is not as common but may have application as a feedback methodology is appraisal by peer group members. Data collected from different appraisers will be compared in the second section. Discussion of the issue of how subjective appraisals are done will center around survey and interview techniques, and this paper will conclude with suggestions for optimizing combined use of these approaches.

TYPES OF APPRAISALS

As outlined above, the types of appraisal of interest here, in terms of what is being appraised, include estimates of task proficiency, criticality, difficulty, and performance frequency. The data summarized below are relevant to the accuracy of such estimates and were selected in accordance with two criteria; they were obtained for specific military tasks or tasks similar to those performed in the military, and they were compared to more objective data obtained in the same study. In many cases in the literature, the accuracy of subjective ratings has been assessed by comparing them to other ratings. Such studies are de-emphasized here in favor of those employing independent objective criteria. At the end of this section, research from the cognitive psychology literature relating to humans' ability to make accurate subjective appraisals is tied in, as appropriate.

Proficiency Appraisals

A key element of feedback from field units to TRADOC Centers/Schools is data relating to the proficiency with which soldiers can perform specific required tasks. Such data are needed to allow elements of Centers/Schools to evaluate both institutional training and unit training and to make modifications as needed. Since the operational testing of soldiers' performance in the field is costly in terms of time and resources, proficiency data are usually gathered through subjective estimates. That is, soldiers are asked to estimate their confidence or the likelihood that they can perform specific tasks. Supervisors may also be asked to rate soldiers' proficiencies. How accurately do such subjective appraisals reflect actual task proficiencies? The relevant data summarized below provide a mixed answer.

Pourchot and Lanning (1979) found that subjective proficiency estimates correlate highly with performance test results in certain instances. Over 200 subjects rated their ability to use hand tools, a task of high relevance to military jobs. These predictions correlated significantly with scores on a performance oriented maintenance test. The authors concluded that the accuracy of the performance appraisals was due to the explicitness of tasks involving hand tools. This suggests that subjective proficiency appraisals can provide accurate performance feedback if the tasks rated are made sufficiently explicit.

Another task of some relevance to the military for which the accuracy of subjective appraisals has been examined is clerical and typing abilities. Levine, Flory, and Ash (1977) found significant positive correlations between subjects' ratings of their abilities and written test scores in areas such as spelling, grammar, reading, and arithmetic. They also found that self-ratings of typing speed correlated at the .60 or higher level with results of a standardized typing test. Ash (1980) further examined the accuracy of self-appraisals of typing ability and found that such ratings correlate moderately well with typing test scores. With a sample of over 150 high school students, self-ratings of straight copy typing ability correlated in the .44 to .59 range with typing tests for alphabetic material, but less than .30 with tests for numeric and tabular material. There was also a lack of discriminant validity in this study. That is, self-estimates of straight copy net words per minute correlated highly with test results for typing of straight copy, letters, and revised manuscripts, but self-estimates of ability to type letters, manuscripts, and numbers did not correlate highly with corresponding test results. Subjects thus demonstrated an ability to accurately appraise their basic straight copy typing speed and accuracy, but they did not accurately appraise more advanced typing abilities with which they had less experience. A leniency error was also found in this study, since the mean straight copy self-appraisal score was approximately 12 net words per minute higher than the mean straight copy test score. A final finding of interest was that minority group members' self-appraisals of straight copy typing abilities were less accurate predictors of test scores than were majority group members' appraisals. The primary conclusion to be drawn from these clerical and typing studies is that subjects can appraise their own abilities with moderate accuracy, as long as the tasks appraised are basic ones with which the subjects have had extensive experience. Secondary conclusions are that leniency errors may occur with such appraisals, and that minority group members may appraise their abilities less accurately.

Within the field of education a large body of research has been reported which relates to the accuracy of subjective appraisals of proficiency. Much of this research has limited relevance to the present review, since it addresses appraisals of general knowledge obtained in a classroom rather than appraisals of task-specific performance abilities. The problem of obtaining an objective criterion to compare subjective appraisals against is exacerbated when one is addressing general cognitive abilities rather than "hands-on" or motor abilities. But despite this criterion problem, educational research has provided some findings of relevance in a military context, particularly since much military training is conducted in a classroom and military skills are becoming more cognitively oriented. Thus, educational research on subjective evaluation or appraisal is selectively reviewed below.

Numerous studies have shown that at least some students can accurately self-appraise their course performance. Moreland, Miller, and Laucka (1981) found that good students were accurate in their self-appraisals, but poor students were relatively inaccurate. The poor students understood the course grading criteria, but for some reason they failed to accurately apply these criteria to their own course work. Shaughnessy (1979) found a similar result by obtaining confidence judgments along with answers to multiple-choice questions. Confidence judgments were found to be moderately accurate, and there was a strong positive relationship between confidence judgment accuracy and test performance. Students who knew an answer knew that they knew. Cohen (1981) reviewed the results of 14 studies in this area and found that the mean correlation between self-appraisals and student achievement on tests was .47. Students are at least moderately accurate in appraising their performance on written tests, and good students are relatively more accurate than poor students.

There is some evidence that teachers are not as accurate in subjectively appraising classroom activities as students are. Hook and Rosenshine (1979) found that teachers' perceptions of classroom activities were inaccurate compared with perceptions of students and outside observers. For example, teachers were found to be inaccurate in appraising the amounts of recitation, discussion, and question answering that occurred in their classes. Teachers' global ratings of classroom activity were found to be moderately accurate compared with observers' ratings, but teachers' appraisals of specific activities were found to be inaccurate. Hook and Rosenshine (1979) concluded that teachers' appraisals of specific classroom activities should not be assumed to correspond to actual practice. Shavelson and Dempsey-Atwood (1976) reached a similar conclusion in a review of the relationships between teacher behavior and student outcome measures. Measures of teacher behavior, including subjective appraisals, were found to be unstable and inaccurate, with global ratings showing the most stability. The appropriate overall conclusion from this line of research is that teachers' appraisals of their specific classroom proficiencies do not agree with outside observers' appraisals. Whether one concludes that teachers are inaccurate or observers are inaccurate (or both), this research provides evidence of the inaccuracy of subjective appraisals.

Cohen (1981) performed a meta-analysis of studies of the relationship between student ratings of instruction and student achievement and found stronger support for the accuracy of student ratings than had previously been published.

The average correlation between overall course ratings by students and student achievement on written tests was .47, and the average correlation between ratings of instructors and achievement was .43. This again supports the accuracy of global ratings, although ratings of somewhat more specific areas such as instructors' skill and course organization were also found to be accurate. Three general variables which influenced the accuracy of course appraisals were identified. Appraisals were more accurate for courses taught by experienced instructors rather than graduate students, for courses in which achievement tests were not graded by students' own instructors, and for courses in which students gave their appraisals after they knew their final grades. The finding of increased accuracy with the use of external graders could be attributed to inconsistencies in grading practices among instructors. Such inconsistencies would lessen the accuracy of subjective appraisals since they would result in an unreliable criterion. The finding that students' appraisals are more accurate when they know their final grade may indicate that teachers can buy good evaluations by giving good grades. If students tend to evaluate what they have learned based on what grade they have achieved, then the accuracy of evaluations would be more appropriately measured in situations where students do not know their final grades. In such studies the correlation between course appraisals and achievement was found to be .38, indicating at best moderate accuracy.

Cohen's (1981) conclusion that students' appraisals of instruction are an accurate index of students' proficiencies (i.e., what they learned from the course) must be tempered in several respects. Most of the appraisals addressed were global in nature and there are indications that students use global factors such as the final grade achieved or expected in evaluating a course or an instructor. Accuracy of global judgments may not be indicative of accuracy in the types of task-specific performance of interest in the present paper. The criterion used in studies of students' appraisals has most commonly been achievement on a written test. Results from such studies may or may not generalize to military situations in which the criterion is manual performance of a task. And, as pointed out by Cohen (1981), achievement on a retention test given at a later time may be a more valid criterion against which to compare subjective appraisals than within-course tests are.

Educational research on the abilities of students and teachers to accurately appraise their course proficiencies has provided somewhat mixed results. But there are several indications that good students can accurately judge what they have learned, in at least a global sense. Further research is needed to determine if this result generalizes to a military context. Such research should address specific tasks and use results of both immediate and delayed performance tests as the criterion.

DeNisi and Shaw (1977) noted that subjective proficiency appraisals addressed in previous educational and other research had generally dealt with broadly defined or global abilities. They attempted to remedy this situation by examining the accuracy of self-appraisals for more specific abilities, such as visual pursuit, manual speed and accuracy, and spatial orientation. College students used five-point scales to self-appraise their abilities on specific tasks and were then tested on each task using ability tests commonly used in industrial settings. Sample test items were used to insure that each student

understood the specific abilities being appraised. Results showed that while correlations between self-appraised and tested abilities were almost all statistically significant, they were too small to be of any practical significance. This finding demonstrates a problem with interpretation of studies of subjective appraisal accuracy, many of which involve correlational analyses. While DeNisi and Shaw (1977) considered correlations in the .20 to .40 range to be of little practical significance, other researchers interpret such correlations as indicating at least moderate accuracy of subjective appraisals (Cohen, 1981). DeNisi and Shaw (1977) supported their interpretation by showing that the self-appraisals failed to differentiate between students who subsequently scored low or high on corresponding ability tests. That is, the predicted test score for students rating themselves high in a given ability was within the 95 percent confidence interval established around the predicted score for students rating themselves relatively low (no one rated themselves below average, indicating a leniency bias). The appropriate conclusion to be drawn from this study is thus that self-appraisals are not sufficiently accurate to be substituted for tests of specific abilities. The practical significance of correlations with a magnitude of approximately .40 is a matter for debate. In line with DeNisi and Shaw (1977), such correlations will not be interpreted in the present paper as strongly supporting the accuracy of subjective appraisals.

The research reviewed thus far in this section has dealt with general knowledge or basic skills which were not appraised in a military setting. In a study of more direct relevance to the Army, Gilbert and Downey (1978) looked at the correlation between 10 measures of performance taken during Ranger training and criterion measures obtained for the same group of officers three years later. Unfortunately, this study did not provide a particularly useful evaluation of the accuracy of subjective appraisal, since both the original and subsequent sets of measures consisted largely of ratings by peers and superiors. Correlations between these two sets of ratings ranged from .11 to .35, indicating a lack of agreement over time, perhaps due to the use of two different sets of raters (low interrater reliability). A halo effect may also have been present, as ratings of individuals on 10 dimensions tended to be highly similar. The validity or accuracy of the ratings could not be determined due to the lack of an independent objective criterion, but the problems described above (low interrater reliability and halo effect) and the fact that the components of performance and their relative contribution to proficiency changed with experience would necessarily limit validity coefficients.

In a study conducted for the US Navy, Hall, Denton, and Zajkowski (1978) used achievement on a job knowledge test as a criterion for determining the accuracy of subjective appraisals of proficiency. During a structured interview, supervisors estimated the proficiency of 32 electricians and boiler technicians on specific tasks. These estimates were compared to the sailors' performance on written tests, and correlations were found to be low and nonsignificant. The authors concluded that interview and written test methods did not produce equivalent information about task proficiency. Comparison of proficiency estimates with "hands-on" performance would have allowed more definitive conclusions about the accuracy of subjective appraisals.

In another study of direct relevance to the Army, Medlin and Thompson (1980) attempted to determine the major dimensions or factors that military judges use in subjectively appraising ARTEP performance. A complex multidimensional analysis of ratings based upon written narratives of ARTEP performance indicated that military judges use only one general rating dimension, indicating a possible halo effect. A general impression of unit performance apparently is used to evaluate the unit, and more specific factors are used only if no strong overall impression is made. Again, the accuracy of subjective ARTEP evaluations could not be determined due to the lack of an independent objective criterion in this study, but appraisals of specific aspects of unit performance could not be expected to be accurate if they are based only upon general impressions.

Caution should be applied in generalizing from the results of this last study, since the appraisals were based upon brief written narratives and not upon actual observation of field performance. But it and the previous studies do demonstrate some important points about many studies of subjective appraisal in a military setting. In many cases an objective criterion is not available to allow full determination of the accuracy of subjective judgments. Ratings are often compared with other ratings. But problems such as low reliabilities and halo effects limit the accuracy that should be expected. The tasks for which performance is subjectively appraised are also often not very specific or explicit, again leading one to expect low judgmental accuracy. Summarized below are studies which avoid these limitations by addressing task-specific appraisals compared with objective performance measures.

Schendel and Hagman (in press) have reported at least indirect evidence for the accuracy of task-specific subjective proficiency appraisals. Soldiers were trained to assemble/disassemble the M60 machinegun and were then retention tested and retrained several weeks later. Before they were retention tested, soldiers were asked to estimate how much refresher training they would require to regain proficiency on the task. These subjective estimates were highly accurate. However, this result does not provide strong evidence for the accuracy of subjective proficiency appraisals, due to the fact that limited retraining was needed. An average of only two trials were required for retraining, and soldiers knew from initial training experience that they would be shown the correct procedure if they made an error during the first trial. It is thus not surprising that soldiers were able to correctly estimate that they could relearn the task within two trials. The accuracy of refresher training estimates should be further addressed using tasks that require large numbers of retraining trials.

Hiller (1980) developed algebraic models for determining the relative benefits (in terms of time saved or lost) of alternative pretesting procedures; i.e., ways of determining whether a soldier needs training on a specific task. The alternative procedures analyzed included self-estimates of task proficiency, written tests, and performance tests. While the original paper did not directly address the relative accuracy of these appraisal methods, Hiller (Note 2) has provided data which allow comparison of self-estimates and performance test results for five specific tasks. Two of these tasks (organize and employ a tank hunter-killer team) involve leadership skills, two (encode/decode and

authenticate messages with a KAL 16 Coding Device) are primarily cognitive in nature, and one (emplace/recover an M16A1 Anti-Personnel Mine) involves "hands-on" motor skills. Self-estimates of proficiency were highly accurate for the two leadership tasks; nearly everyone who said they could do each task passed the performance test, and everyone who said they could not do each task failed the performance test. But cognitive tasks showed considerably less accuracy in self-appraisals; only 46% of those who said they could authenticate a message could actually do so, while 50% who felt they could not do the task passed the performance test. Corresponding results for encoding/decoding messages were 37% and 25%. Finally, accuracy of self-estimates was especially low for "hands-on" skills; only 23% of soldiers who said they could emplace and recover a mine could actually do so, while 32% of those who said they could not do this task were able to pass the performance test. So the accuracy of subjective appraisal in this study depended upon the type of task being addressed. Why did this occur? One possible reason is that the accuracy of subjective appraisal declines as the criterion with which it is compared becomes more objective. Leadership skills are difficult to develop standards for and objectively evaluate; the high accuracy for self-appraisal of leadership skills described above may have resulted from the comparison of two subjective appraisals. That is, the performance tests for the two leadership tasks may have been relatively subjective in nature. The performance test standards for the cognitive skills would be expected to be more objective, resulting in less accuracy of subjective appraisals. And the test standards should be the most objective for the "hands-on" task, which showed the least subjective appraisal accuracy. This interpretation of the results indicates that subjective self-appraisal of proficiency on specific tasks is not accurate when compared with an objective criterion. This conclusion is admittedly based upon a small sample of tasks, so further relevant data are summarized below.

Shields, Goldberg, and Dressel (1979) examined the retention of 20 basic soldiering skills by administering performance tests to soldiers in the field. The tasks addressed included such basic skills as first aid, challenge and password, donning the gas mask, and checking the field telephone. As a part of this study, confidence ratings of proficiency (self-appraisals) were obtained using a four-point scale for each task before it was tested. While the report referenced above does not directly discuss the relationships between confidence ratings and task performance, some indication of inaccuracies in self-appraisals can be gleaned from it. For example, 75% of the confidence ratings collected indicated that a task could be performed fairly well or very well, but only 37% of the tasks were correctly performed with no coaching during the tests. This may be an indication of leniency errors. Goldberg (Note 3) has provided further analyses of the results of this study, and the relationship between confidence and performance was found to be consistently low. Correlations examined for several tasks ranged from -.30 to .06. Goldberg (Note 3) has also reported that later studies of retention of artillery skills showed a similar lack of correlation between confidence judgments and task performance. Correlations in the .40 to .50 range were found between averaged confidence ratings and averaged performance scores, perhaps indicating some ability to accurately appraise performance in general, but consistently low correlations were found between confidence and performance on specific tasks. It is interesting to note that the non-relationships described above have not

been discussed in published reports. Other retention studies (e.g., Rose and Wheaton, 1978) have been found in which subjective appraisals of proficiency were collected but their relationship to performance was not reported. It is probably a safe conclusion that no significant relationships were found in such studies, and that retention research in general has not found subjective appraisals of proficiency to be accurate.

In summary, the data reviewed above indicate that subjective appraisals of proficiencies (largely in terms of self-appraisals) on specific tasks often do not represent true abilities. This appears to be especially true when the subjective appraisals are compared to objective well-specified performance criteria. If subjective appraisals are influenced by leniency errors (the data above indicate that they are), and if the performance criteria are also subjective and lenient, then a falsely high relationship can be expected between these two measures. Before subjective appraisals are used as feedback from field units to Centers/Schools, the relationship between such appraisals and more objective measures of performance should be further examined. Such examination should use task-specific performance tests with valid objective standards. Self-ratings of proficiency may only be accurate when addressing explicit tasks with which the ratees have extensive experience. This point will be further addressed in a later discussion of ways to improve the utility of subjective appraisals.

Task Criticality Appraisals

Another type of subjective appraisal of concern to TRADOC Centers/Schools is estimation of task criticality. Limited resources and time do not allow training of all tasks in a given MOS in the training institution. Training developers must thus somehow decide which tasks are most critical for combat performance and therefore most important to train. This is typically accomplished by preparing an extensive list of tasks and asking subject matter experts to subjectively rate their criticality, usually by employing some sort of rating scale. These experts may be drawn from personnel available in the training institution, or feedback may be solicited from personnel in field units (often through CODAP surveys). In either case, the judgments are based upon field experience and thus represent subjective feedback from the field to Centers/Schools. Just as with estimates of proficiency, one can question how accurately subjective appraisals of criticality represent the "true" relative importance of tasks.

Data are relatively sparse in this area, but those available have been summarized by Harris, Osborn, and Boldovici (1978). These authors conclude that a key problem with criticality estimates is that rater agreement (inter-rater reliability) has generally been found to be low. They also conclude that nothing is known about the predictive validity of criticality ratings, or the degree to which such ratings correlate with more objective measures of task criticality (of course, one of the problems here is developing objective measures of criticality). Since such measures cannot be developed during actual combat, they must be developed using simulations and war games, which can be costly and time-consuming. But as long as the reliability of criticality ratings is low, their accuracy or predictive validity also will be low. Harris,

Osborn, and Boldovici (1978) suggest several ways in which the reliability of criticality estimates can be increased, such as using paired-comparison techniques for determining the relative rather than the absolute criticality of tasks. These techniques will be addressed in a discussion of ways to improve the accuracy of subjective appraisals in a later section of this paper. The important point for now is that the relevant data available do not suggest that subjective appraisals of task criticality are reliable or accurate. If accurate measures of task criticality are desired, further work is needed to make criticality ratings more reliable and objective.

Task Difficulty Appraisals

The next type of subjective appraisal to be discussed here involves judgments of the difficulties of tasks. Such appraisals are important to Centers/Schools since the relative difficulty of tasks influences the distribution of training time and resources. If particular tasks are more difficult for soldiers to perform and retain, they should be given increased emphasis in the training base or retrained more often in units. Appraisals of task difficulty are often made subjectively; that is, training developers decide, based upon their experiences and the opinions of subject matter experts, how training resources should be distributed across tasks. How accurate are experts' appraisals of task difficulty? The two sets of relevant data summarized below indicate that the accuracy may be rather low.

Ryan-Jones (1979) obtained squad leaders' and platoon leaders' ratings of difficulty for 18 basic infantry tasks and compared them with the percentage of soldiers failing each task on the written component of the Skill Qualification Test (SQT). The correlation between these two sets of measures was low (-.38), indicating that experts' ratings of difficulty may not be representative of actual task difficulty. This interpretation is based on the assumption that the written component of the SQT is representative of actual task performance. If this assumption were not correct, one could conclude that the experts were right but the SQT is wrong. What is needed is a comparison of experts' ratings with actual hands-on performance results. Harris, Campbell, and Osborn (1979) accomplished this by comparing expert ratings obtained from training developers and senior NCO's with performance results obtained during the Army Training Study (ARTS; 1978). The experts' difficulty ratings were found to be unreliable and unrepresentative of performance. For example, when experts were asked to select the most difficult element of a task, they selected the element most often performed wrong only 16% of the time. Using a more lenient criterion, they selected one of the three most commonly failed elements of a task only 45% of the time. Thus, indications are that subject matter experts are not accurate in appraising the difficulty of performing tasks or elements within tasks. It may be that experts' conceptions of tasks differ from those of novices, leading experts to be unable to predict where relative novices will encounter difficulties. In any case, experts' ratings of task difficulty should not be accepted as accurate without further comparison with objective performance data.

One possible reason for the lack of reliability and accuracy that has been found in ratings of the difficulty of tasks may lie in the way that difficulty

has been subjectively appraised (Hiller, Note 4). When one is asked to judge the difficulty of a task, one can interpret and answer the question in various ways. The task may be difficult to train or teach, difficult to learn, or difficult to perform once learned. These differing interpretations of difficulty will not always lead to the same subjective judgments. For example, learning to encode/decode and authenticate messages is fairly difficult, but these tasks are easy to perform after they are learned. Conversely, learning how to locate an anti-personnel mine is easy, but performance of the task is painstaking, stressful, and difficult. If subject matter experts rating tasks such as these differ in their interpretation of whether they are judging learning or performance difficulty, their ratings will not agree and interrater reliability will suffer. Thus, when appraisals of task difficulty are obtained, the difficulty dimension should be operationally defined in terms of teaching, learning, or performing. In this way the reliability and accuracy of these appraisals can perhaps be increased. This hypothesis is supported by Hiller (1974), who found that students' ratings of text readability (difficulty) corresponded to objective measures of comprehension on both immediate and delayed retention tests. The accuracy of these appraisals may have been due to the definition of difficulty in terms of a dimension (readability) for which the raters shared a common understanding.

Task Frequency Estimates

Developers of training programs may need to know how frequently specific tasks are performed in the field, in addition to how critical and difficult they are. Tasks which are performed frequently generally require less sustainment training. Tasks which are not performed frequently may be important ones to include in unit training. If an infrequently performed task is also a critical one for combat performance, a unit training program should be developed for it in order to lessen retention problems. So frequency considerations can interact with those of criticality and difficulty.

There are few data available relating to the accuracy of subjective task frequency judgments. Various studies of skill retention (e.g., ARTS, 1978; Rose and Wheaton, 1978) have obtained such judgments from soldiers in the field in order to examine the effects of practice upon retention. Little relationship has typically been found between these two variables, which may indicate that no relationship exists, or that the frequency estimates obtained have not been accurate. Turney and Cohen (1978) also obtained data of relevance by comparing self-estimates of work effort and time with actual performance duration for three tasks in a computer facility. The correlations of estimates and actual effort were in the .30 to .40 range, indicating only moderate accuracy in self-appraisals of time and effort expended.

It is very difficult to obtain objective measures of task performance frequency, since one would be required to observe the activities of individuals in a unit and count task performances over a long period of time. Unit records are generally not detailed enough to provide task performance frequency counts. Job books might be expected to provide such data, but they are often incomplete and difficult to consolidate (Burnside, 1981).

Only one study has been identified which directly compared subjective estimates of task performance time and frequency with observed performance in a field setting. Johnson, Tokunaga, and Hiller (1980) reviewed the available literature and concluded that objective methods were needed to validate self-appraisals of time spent performing specific tasks, since previous studies indicated that such appraisals were not likely to be accurate. They then asked a sample of 98 officers and NCO's in Infantry companies and Artillery batteries how often they performed each of a large set of tasks in a typical month, and how long it took to perform each task once. These two estimates were combined by the researchers to obtain absolute estimates of the time spent on each task in a typical month. These estimates were compared with data obtained by observing the activities of 56 personnel within their units. Personnel were observed for an average of about four hours each, and the dominant behavior within each ten minute interval was recorded. The tasks addressed in the subjective estimates of frequency and time spent were categorized into broad content areas for comparison with the observational data. The rank order correlations between subjective estimates and observational data were found to range from .65 to .90 for various levels of personnel, indicating that the estimates were highly accurate. The estimates were found to inflate the absolute amount of time spent at work, but they were reliably related to the observation criterion. Converting the time estimates to proportions by dividing them by the total time estimates yielded a truer picture of the distribution of time across tasks.

Why did Johnson, Tokunaga, and Hiller (1980) find that subjective estimates of time spent performing tasks were accurate when this result has not been found elsewhere? Two possible reasons can be identified. First, the comparison of time estimates and observational data was accomplished in terms of broad categories of tasks, and not for specific tasks. It may be that time estimates are more accurate for general tasks than for specific tasks. Further research with precise observational data would be necessary to determine if this is the case. Secondly, Johnson, Tokunaga, and Hiller (1980) broke the time spent estimates down into two estimates, one for how often a task is performed and one for how long a typical performance takes. These two estimates may be relatively simple to give and thus relatively accurate. If this is the case, we have evidence that frequency estimates can be relatively accurate and that subjective estimates in general can be made more accurate by asking more precise questions. More research using objective observational criteria is needed to further address these indications.

Appraisal of Training Materials

All the types of subjective appraisal discussed above are related to some aspect of performance on specific tasks. TRADOC Centers/Schools also have a mission to appraise the quality of individual and collective training materials they produce, such as Soldiers' Manuals, ARTEP's, commanders' guides, and crew drills. The appraisal of these materials is also accomplished largely through subjective approaches, such as the receipt of informal comments and the administration of questionnaires (Burnside, 1981). The issues addressed for materials are similar to those addressed for task performance, such as the criticality of the information in the documents, the frequency of documents'

use, and the degree to which they enhance mission performance. One can address the accuracy of subjective appraisals of these issues for associated training materials as well as for task performance, although little research has been done in this area.

One study of relevance (Shvern, 1979) examined evaluations of a combat commander's guide obtained via a questionnaire. There was an indication that sections of the guide were not evaluated independently, since they tended to be rated the same. This is evidence of a halo effect, similar to those described earlier. Another finding was that each rating depended largely upon the unique measure used and its context, making generalization difficult. Some of the problems encountered in subjective appraisals of task performance may also occur in subjective appraisals of materials. Conclusions and suggestions offered in this paper should thus be applied to both areas of evaluation.

Tentative Conclusions

What can one conclude about the accuracy of various types of subjective appraisal? One appropriate conclusion is that directly relevant data are scarce. Few studies have gathered comparative data using an objective criterion in order to directly analyze the accuracy or validity of subjective data. But studies which do allow such comparisons, as well as studies of other aspects of subjective appraisals (e.g., reliability and halo effects), indicate that subjective data are often inaccurate. There is some indication that subjective appraisals may be at least moderately accurate when they address explicit tasks with which the appraiser has extensive experience. But there is also some indication that subjective appraisals become less accurate as they are compared to more objective criteria. And there is evidence of the types of errors discussed in the first section of this paper in subjective appraisals gathered in a military setting. Raters tend to disagree with each other (low interrater reliability), tend to make general judgments without distinguishing among the different aspects of a situation (halo effect), and tend to provide positively biased ratings (leniency error). Obviously, further research is needed to identify the extent of such problems in subjective appraisals, and to identify ways of reducing or eliminating them. Initial steps in this direction are discussed in a later section of this paper.

Cognitive Psychology

Subjective self-appraisal or the estimation of one's own abilities to perform specific tasks would likely be classified as introspection in the experimental psychology literature. Introspection involves the observation by a person of his or her thoughts and feelings and verbal reports or behavior describing them. This technique was widely utilized during the early days of experimental psychology, but was abandoned following behaviorism's emphasis on the analysis of objective behavior. However, rebirth of interest in the study of unobservable mental processes within cognitive psychology during the past twenty years has led to a reemergence of research on the accuracy of introspective reports. Most of this research has been directed toward introspections of higher cognitive processes such as problem-solving, but it may have some relevance to introspections of task-specific abilities.

Lieberman (1979) has issued a call for a limited return to introspection as an experimental technique, since it may be accurate in some instances. For example, people are able to accurately appraise and state how they will vote, as shown by the accuracy of polls. There are several examples of accurate subjective appraisals in the cognitive psychology literature. Carver (1972) reported that subjective estimates of the percent of thoughts understood during reading correlated .98 with a test measuring the amount of information stored. This finding demonstrates an ability to subjectively appraise the difficulty of a highly familiar task such as reading. Kroll and Kellicutt (1972) showed that people were able to accurately predict how well they could recall verbal material by reporting how many times they had implicitly rehearsed it. Lachman, Lachman, and Thronesbery (1979) found that people who couldn't recall the answers to general knowledge questions were able to accurately predict whether they would recognize the correct answers. They also were found to spend more time searching memory for answers they thought they knew than for answers they thought they did not know, which perhaps led to a self-fulfilling prophecy. Both Robinson and Kulp (1970) and Gardiner and Klee (1976) found that people are able to accurately recognize most of the items from a verbal list that they previously recalled on a free recall test.

The evidence summarized above indicates that people can accurately appraise their past and future memory abilities, at least when familiar verbal material is involved. This higher-level knowledge of memory abilities has been christened metamemory (Flavell, 1970). Metamemory has been shown to be accurate for general knowledge and frequently used memory abilities, and for episodic (Tulving, 1972) tasks such as recall or recognition of verbal items presented in lists. Is metamemory available and accurate for complex motor skills which may not have been practiced extensively? Metamemory for specific motor abilities may be available only in a general sense. That is, soldiers might know that they had performed a task before and be able to verbally describe its general characteristics, but still be unable to accurately appraise whether they can perform the task, due to forgotten details or misunderstood standards. The characteristics of metamemory for complex skills and the extent to which accurate introspections can be derived from it are important topics for future research. As pointed out by Lieberman (1979), introspection should not be totally rejected as an inaccurate technique, but rather the conditions under which it is likely to be accurate and useful should be identified. In order to do this, introspective reports should be supplemented and verified by other behavioral or circumstantial evidence, whenever possible.

While arguments for the use of introspection in some instances certainly have merit, the accuracy of this technique is still a subject of debate in the cognitive psychology literature. Kahneman and Tversky (1973) have argued that subjective judgments and predictions are based upon general heuristics rather than upon specific evidence available. Their research shows that one predicts by selecting the outcome that is most representative of the input, even when this outcome is statistically unlikely. For example, subjects were asked to predict the major area of study for a particular student, based upon a written personality description. When the personality description was stereotypical of that for an engineer, subjects predicted that the student was an engineering major. They persisted in this prediction, even when told that the frequency

of engineering students was very low and that the personality description might not be accurate. Kahneman and Tversky (1973) concluded that prior probabilities are ignored when stereotypical evidence is available, even if that evidence is worthless.

Extrapolating from the findings described above to the sorts of task-specific self-appraisals of interest in the present paper, it may be that soldiers estimate their proficiencies in terms of what they should be able to do rather than in terms of what they can actually do. That is, if a soldier is asked whether he can properly perform a particular task, he may respond positively because he feels that a soldier with his level of experience should be able to perform it. He may not have actually thought out whether or how he could perform the task. The soldier may respond on the basis of a stereotype or implicit theory about the abilities of soldiers at his level. Nisbett and Wilson (1977) have supported such a contention with research showing that people do not base reports of their cognitive processes on true introspections. Rather, their reports are based on implicit causal theories about the extent to which particular stimuli are plausible causes of specific responses. They describe introspection as nothing more than judgments of plausibility and conclude that "the accuracy of subjective reports is so poor as to suggest that any introspective access that may exist is not sufficient to produce generally correct or reliable reports" (p. 233). Accurate subjective reports would then only occur incidentally as the result of use of a correct implicit theory about behavior. Such reports could not be expected to be generally accurate if people cannot introspect about their mental processes. But this is not the end of the matter. Smith and Miller (1978) have challenged Nisbett and Wilson's (1977) conclusions on theoretical and methodological grounds, and they have argued that people can accurately introspect about their mental processes in some instances. These instances include tasks which are novel, engaging, and not overlearned, so that the mental processes involved are not automatic and unconscious. These authors suggest that research be oriented not on the question of whether people can introspect about mental processes, but rather on the question of the conditions under which such introspection is accurate.

In summary, what does the cognitive psychology literature offer that has relevance to the sorts of subjective appraisal of interest here? First, a caveat mentioned above should be repeated. Research on subjective judgment within cognitive psychology has primarily addressed higher mental processes. Findings in this context may or may not directly relate to judgments about abilities which are more motor or "hands-on" in nature. However, many of today's military tasks are cognitively oriented, so findings from the cognitive research literature should have some application in a military setting. Analyses of the accuracy of subjective judgments in cognitive settings have produced mixed results and have not yet provided convincing evidence that such judgments are accurate. Lieberman (1979) and Smith and Miller (1978) have suggested that debates about the general accuracy of subjective judgments should be replaced by research addressing the conditions under which such judgments can be accurate. The present paper will attempt to encourage movement in this direction by describing ways in which subjective appraisals may be made more accurate. The military and cognitive research literature will be integrated in the development of these suggestions after review of findings concerning types of appraisers and appraisal methods.

TYPES OF APPRAISERS

A primary consideration in the use of subjective appraisals is the sources from which they are collected. In situations such as the gathering of subjective appraisals as feedback from military units in the field, three general alternative sources are available: soldiers evaluating themselves (self-appraisal), supervisors, and peer group members. For example, suppose that Center/School personnel wish to economically appraise soldiers' proficiencies on specific tasks. Soldiers could be asked to subjectively appraise their own performance on the tasks, supervisors could be asked to appraise the performance of soldiers working under them, or soldiers could be asked to appraise the performance of their co-workers or peers. A previous review indicates that the first two of these alternatives are the ones most commonly utilized by TRADOC Centers/Schools (Burnside, 1981). The previous section of the present paper summarized data relevant to the absolute accuracy of subjective appraisals. This section summarizes data relating to the relative accuracy of appraisals obtained from alternative sources, particularly supervisor versus self-appraisals.

What are the relative plusses and minuses in utilizing self-appraisals versus subjective appraisals gathered from other sources? A primary benefit of self-appraisals pointed out by numerous authors (e.g., Levine, 1980; Primoff, 1980; Shrauger and Osberg, 1981) is that individuals have extensive data available about themselves and can provide information that is unavailable from other sources. We observe ourselves continuously in our daily work settings, while supervisors and peers may have limited opportunities to observe our performance. Given basic self-observation and memory capabilities, we should then have more information available relating to our abilities than any other source. However, a note of caution is appropriate here. Recall that some of the cognitive psychology literature summarized earlier (e.g., Nisbett and Wilson, 1977) calls into question our ability to introspect about our own capabilities, at least those that are cognitive in nature. But until this issue is resolved, we can at least theoretically expect self-appraisals to benefit from the relatively large amount of information available. A related potential advantage of self-appraisals is that individuals generally attend to situational factors in their own behavior, whereas outside observers may not be aware of such factors (Wills, 1978; Shrauger and Osberg, 1981). Individuals might thus be expected to be more accurate in appraisals of their own abilities, since outside observers might tend to over-generalize across situations. In fact, Wills (1978) has shown that observers tend to regard small samples of others' behavior as sufficient evidence for generalized personality dispositions. Supervisors and peers may similarly tend to over-generalize about abilities based upon a small sample of data. A final more practical advantage of the self-appraisal approach is that it is likely to be more economical, in terms of time and resources, than are other approaches.

One major disadvantage of self-appraisals is that alluded to above; i.e., people may not be capable of appraising themselves competently. We may not be aware of many of our cognitive and motor abilities, since some of them may be automatic and unconscious. Further basic research will be necessary to resolve this concern; thus far, research and theories relating to our ability to

evaluate task-specific proficiencies have been virtually nonexistent. The second major concern with the use of self-appraisals is the possibility of response biases. We may have more information available about ourselves than anyone else has, but we also have more reasons to bias our appraisals in a positive direction. This would result in a leniency error of the sort described earlier.

Relative Accuracy of Self-Appraisals

Shrauger and Osberg (1981) have examined the utility of obtaining self-appraisals and appraisals from other sources in a variety of situations. Since some of these situations have at least indirect relevance to the military, the review's conclusions will be summarized here. In the area of academic achievement, self-appraisals were found to predict academic performance at least as well as most projective tests that have been utilized. But self-appraisals did not do as well when compared with previous performance in the same situation. That is, college grades were better predictors of future college grades than self-predictions were. Self-appraisals did show higher predictive accuracy than performance in a previous situation; i.e., self-appraisals predicted college grades better than high school grades did. This leads to the conclusion that self-appraisals may be useful when performance indicators gathered in the situation of concern are not available. Self-appraisals of task proficiencies may be accurate relative to results of written knowledge tests, but not relative to results of actual "hands-on" performance.

With respect to the use of self-appraisals to predict actual job performance, Shrauger and Osberg (1981) found few data available in settings other than the Peace Corps. And the results from this setting were not found to be particularly useful, since they were not consistent and involved comparison of self-appraisals with appraisals by peers and supervisors, rather than with more objective measures of on-the-job performance. Conclusions reached in this area were that sufficient data are not available to determine how well people can appraise their performance relative to appraisals developed by evaluation boards of supervisors and peers, and that surprisingly few data are available, in general, to address the usefulness of self-predictions of job performance.

After comparing self-appraisals with other methods of prediction in numerous areas, Shrauger and Osberg (1981) found that 29 studies showed self-appraisals to be more accurate, while 10 favored other appraisal methods. This result seems to support the use of self-appraisals, but two caveats are in order. First, the accuracy of self-appraisal was found to vary with the type of behavior being predicted. Self-appraisals did well in general areas such as vocational choice and judgment of personality traits, but were found to be inconsistent in more specific areas such as job performance in the Peace Corps. Second, no adequate comparisons of self-appraisals with objective measures of job performance were found. The predictive accuracy of self-appraisals has been compared with predictions derived from projective tests, evaluation boards, and other general assessment techniques, but it has seldom been compared with objective measures of actual job performance. The conclusion that self-appraisals are as good as other appraisal methods may indicate that all methods are equally poor, and not that self-appraisals are accurate.

One study conducted in a military setting has supported the relative accuracy of self-appraisal techniques, but it also suffers from a weakness discussed above. Dyer and Hilligoss (1979) obtained self-appraisals and other predictors of job performance for over 400 officers and NCO's in an assessment center. The criterion with which these predictions was compared was field leadership performance ratings obtained from superiors, peers, and subordinates of these personnel six to 18 months after assignment to a unit. Again, the criterion is not really objective and what we have is essentially a comparison of two sets of subjective ratings. Results showed that 11 to 14 percent of two types of self-appraisal measures correlated significantly with the criterion, while only nine percent of assessment exercises and seven percent of peer ratings provided successful predictors. This result might be used to argue for the relative accuracy of self-appraisals, but more interesting is the low predictive accuracy of any method. Even when using another subjective measure as the criterion, only a small percentage of self-appraisal measures were found to accurately predict future performance.

Thornton (1980) has provided a thorough review of the accuracy of self-appraisals of job performance using the framework of types of errors or problems discussed earlier in the present paper (i.e., leniency errors and halo effects). This framework will be used here to summarize his conclusions and those of other authors, where appropriate. With regard to leniency errors, many studies have shown that individuals rate themselves higher than they are rated by others. Self-ratings have been shown to be higher than ratings by supervisors, peers, and assessment center raters. Holzbach (1978) also concluded that self-ratings are more lenient than ratings by supervisors or peers, and that supervisor and peer ratings do not differ significantly. Meyer (1980) has summarized years of research which led to the conclusion that most people have an unrealistically positive perception of their own job performance. He found that typically at least 40 percent of employees rate themselves as being in the top ten percent of performers, and that almost no one rates themselves as being below average. He also found that publicly announced self-appraisals tend not to be as positively biased as those given in confidence.

This last finding reported by Meyer (1980) brings up an important point about the accuracy of self-appraisals. Although self-appraisals have generally been found to exhibit leniency errors, this is not always the case (Van Rijn, 1980). Special measures can be taken to reduce the occurrence of such errors. For example, self-appraisals may be less lenient if the rater knows that his or her supervisor may see the ratings. Leniency errors can also be reduced if the rating scale does not require the rater to compare himself or herself to an average task performer. People are hesitant to rate themselves as being "below average," but may be willing to rate themselves as "better than 25 percent of task performers." Ratings may also be less inflated or lenient if they are verifiable (Van Rijn, 1981). The accuracy of at least a sample of any set of obtained self-appraisals should be compared with objective measures of performance, such as a "hands-on" test. If raters are aware that they will be tested on task performance after giving their self-appraisals, they may tend to be more accurate. Both Mitchell (1979) and Frick and Semmel (1978) reported a related finding that observers report the behavior of others more accurately when they know that the accuracy of their observations is being checked. The accuracy of self-appraisals should be checked, if at all possible, in order to reduce leniency errors.

Unless special measures are taken to eliminate them, leniency errors are likely to be a serious problem when using self-appraisals. In fact, the problem may be even more severe than is indicated by the literature (Van Rijn, 1981). In most of the relevant research, self-appraisals have been gathered in experimental settings in which raters know that their self-ratings will have no real effect on aspects of their future job environment, such as promotional opportunity. If self-appraisals were to have a real impact on the job, the tendency for inflation of ratings might become even more evident. In a military setting, soldiers might inflate their self-ratings of task proficiency if they felt that this would in any way increase their opportunity for promotion. They might also inflate self-ratings in order to avoid participating in retraining for tasks they feel they cannot do. The problem of leniency in ratings means that great care should be taken in utilizing self-appraisals in the real world. Measures such as those suggested above should be applied to reduce leniency, but further research is needed to determine the effectiveness of such measures in real-world settings.

Primoff (1979) summarized several sets of data allowing comparison of supervisory and self-appraisals, and he concluded that there may be more random error in supervisory ratings. This appears to be due to supervisors having inadequate opportunity to observe the behavior being appraised. MacLane (1977) operationally defined unreliability of appraisal as an error in which raters gave different ratings to the same ratee for different statements concerning the same dimension. Supervisors demonstrated errors or rating inconsistencies in 27 percent of their appraisals, while the self-appraisal error rate was only nine percent. Supervisors seemed to lack information about the people they were rating, and they were frequently unable to support their appraisals with examples of behavior on the job. Self-raters were able to provide such support; as stated earlier, one advantage of self-ratings is that people have extensive information available about themselves. Self-appraisals may be more accurate than supervisory appraisals in situations where individuals have extensive experience performing the tasks being appraised and supervisors have not had extensive opportunities to observe task performance.

Thornton (1980) found that in the few studies which have reported variance in ratings, most found less variation in self-appraisals than in appraisals from other sources. However, the halo effect has generally been found to be lower for self-ratings. Holzbach (1978) and Van Rijn (1980) have also found that appraisals by supervisors tend to show a greater halo effect than self-appraisals do. This result is probably related to the earlier discussed finding that people tend to be aware of specific situational determinants of their own performance and are thus less willing to over-generalize than external observers are. Halo effects are thus not as large an area of concern for self-appraisals as for subjective appraisals from other sources. The reasons for a reduced halo effect occurring in conjunction with reduced variance in ratings are unclear and need further examination.

In reviewing studies which directly addressed the relative accuracy of various appraisal sources, Thornton (1980) reported finding inconsistent results. Eleven studies showed a lack of agreement between self-appraisals and appraisals from supervisors or peers, while seven studies found at least partial

agreement between rating sources. Other studies have shown that self-ratings are often not reliable or stable, and thus could not be expected to demonstrate validity. These findings suggest that job holders have a different view of their job performance than other people do, and that self-appraisals should be used very carefully. Evidence for the accuracy of self-appraisals is at this point meager (Van Rijn, 1981). Further work is needed to identify those situations in which self-appraisals may be accurate.

Peer Appraisals

The discussion above has centered around the self-appraisal approach, since this is the method most commonly used for gathering subjective appraisals. Another method which has not been frequently used in gathering feedback by Centers/Schools but which deserves further consideration is peer appraisal. The research summarized above indicates that peer appraisals are more similar to supervisor appraisals than they are to self-appraisals, and that the relative accuracy of these different approaches has not adequately been addressed. Reviews of the peer evaluation literature have provided mixed conclusions about the characteristics of this approach. Downey and Duffy (1978) concluded that peer appraisal methods have demonstrated substantial validity and thus provide a useful tool for predicting performance. Lammlein and Borman (1979) found that peer ratings show high interrater agreement and provide good predictions of future performance. They did not provide enough detail on the studies reviewed to indicate how they reached this latter conclusion. Kane and Lawler (1978) reviewed some of the same literature and reported that no studies included an adequately objective measure of performance. The research on accuracy of peer appraisals compared to objective criteria thus appears to be open to differing interpretations. Kane and Lawler (1978) also reported that no studies have allowed a direct comparison of the accuracy of supervisory and peer ratings, while Lammlein and Borman (1979) concluded that ratings from these two sources correlate moderately well. The relative accuracy of peer appraisals is still a subject of debate; reviewers looking for objective criteria have found no reason to conclude that such appraisals are accurate. Peer ratings may have some characteristics (e.g., high interrater agreement) which make their use desirable in feedback systems. However, as with self-appraisals, peer appraisals should be used carefully in conjunction with a check on their accuracy, since their general accuracy has not been consistently demonstrated in the research literature thus far.

Tentative Conclusions

Research on the relative accuracy of subjective appraisals gathered from various types of sources has left many questions unanswered. It is difficult to address the relative accuracy of appraisal sources when the absolute accuracy of each of them is undetermined. What is needed is a study which includes the collection of supervisory, peer, and self-predictions of proficiencies on specific tasks, followed by objective measures of task performance. The literature thus far has generally failed to include objective criteria for comparison purposes, and until it does the accuracy issue will be unresolved. Self-appraisals usually suffer from leniency biases, and peer and supervisory appraisals may suffer from tendencies to over-generalize from small samples of data. Accuracy of these approaches should thus not be assumed, but should be checked against relatively objective criteria.

TYPES OF APPRAISAL METHODS

The final issue to be addressed relates to methods which can be used in collecting subjective appraisals. The data reviewed thus far suggest that subjective appraisals should not be indiscriminately used as feedback to Centers/Schools, since the accuracy of such appraisals is yet to be fully determined. But subjective appraisals are going to be used in the real world, due to the relative ease and economy with which they can be collected. Thus, authors such as Lieberman (1979) and Smith and Miller (1978) are correct in the assertion that it is more fruitful to identify methods and situations which allow one to maximize the reliability and accuracy of subjective judgments, rather than to debate at length the general accuracy of such judgments. In keeping with this suggestion, the remainder of this paper will concentrate upon methods for increasing the accuracy of subjective appraisals. Methods discussed in this section will lead to recommendations and suggestions summarized in the next section.

Surveys and Interviews

Since surveys and interviews are the most commonly used approaches for gathering subjective feedback data, the first issue to be addressed here is which of these methods should be used in specific situations. Survey data have the advantage of being easy and economical to collect, particularly if they are gathered through the mail. However, data summarized by Burnside (1981) indicate that response rates to mailed surveys are often so low as to make this approach to gathering feedback inadequate. In order to gather survey data from a representative sample, it is generally necessary for a data collector to be on-site in the field. The interview approach has the advantage of allowing collection of more in-depth responses, but it is considerably more resource-intensive. Interviews are usually conducted in a one-on-one setting, and this leads to extensive time commitments on the part of data collectors. But this may be time well spent. Burnside (1981) found that battalion staff personnel feel that they give more thoughtful and in-depth answers to interview questions than to survey questions. These personnel are sometimes so inundated with surveys that they do not take time to respond to them carefully, if at all. The use of interviews may thus in some cases result in collection of more valid data.

Hall, Denton, and Zajkowski (1978) conducted a direct comparison of feedback data gathered by mailed questionnaire and structured interview techniques for several tasks in the Navy. Results indicated that these approaches produced equivalent data pertaining to the adequacy of initial training, the frequency of task performance, and supervisors' appraisals of on-the-job proficiency. However, the interview used here was essentially an orally administered survey, so equivalence of results is not surprising. Problems were encountered in obtaining a satisfactory return rate for surveys, demonstrating a common problem with this technique. This study shows that equivalent subjective appraisals can be obtained in response to written or oral questions, if one can get around the problem of low return rate of surveys. But a more interesting issue than how survey and interview responses can be made equivalent is how they can be designed to supplement each other. Surveys can be used to obtain

a general overview of where problem areas lie. Interviews can then be used to obtain more in-depth data on specific problems and the reasons for them. Incidentally, Hall, Denton, and Zajkowski (1978) not only found that survey and interview responses were equivalent, but they also found that proficiency ratings obtained in interviews did not correlate significantly with results of written knowledge tests. When surveys and interviews are used to gather subjective feedback data, a check on the accuracy of such data should be included. A total feedback system should thus use surveys, interviews, and objective tests in conjunction.

Phrasing of Questions

Another important methodological issue in the collection of subjective appraisals is the nature of the questions asked. Meyer (1980) has provided an example of how this variable can influence the value of the information gathered. Self-appraisals which involve the comparison of one's abilities with those of others on specific tasks often lead to leniency errors. But comparison of one's own relative strengths on different tasks may lead to reasonably accurate and useful ratings. Questions should perhaps be phrased to ask self-appraisers to compare their own relative strengths in abilities, rather than to compare their abilities to those of others. When a rating scale requires a respondent to compare his or her performance with the performance of others, the respondent must have knowledge not only of his or her own abilities, but also of others' abilities. Since such scales require an assumption of additional knowledge, they should be avoided where possible.

Bernardin, Beatty, and Jensen (1980) suggested that subjective rating instruments should be based upon a thorough job analysis, and Primoff (1980) provided some further recommendations in this direction. Designers of subjective appraisal questions should be certain that they have an understanding of job elements in common with that of raters. A question designer who is an expert on the tasks addressed may have a different concept of adequate task performance than a rater who is a relative novice. If possible, rating scales should be phrased in terms of explicit behavioral measures of performance rather than in general terms such as "can do the task with no problems." Or, raters could be asked to provide specific experiential evidence supporting their claims that they can perform particular tasks. Appraisals based on observable behaviors are more closely related to task performance than are appraisals based on general factors, such as inferred personality traits (Van Rijn, 1980). A common base or standard for ratings should be ensured between question developers and raters. If raters are asked whether they can perform a task to standard, care should be exercised to ensure that they have the correct standard in mind. Care should also be exercised to ensure that all raters interpret the rating dimension similarly. As described earlier, a general dimension such as task difficulty can be interpreted in various ways, so it should be operationally defined to raters.

Shrauger and Osberg (1981) have recommended ways in which questions can be phrased to maximize accuracy, in addition to the general suggestion that the situation and behavior to be predicted should be specified exactly. There is

some evidence that ratings of maximal behavior result in more accurate predictions of future actions than do ratings of typical behavior. Developers of appraisal questions should be aware of whether the criterion they are interested in involves maximal or typical functioning. Questions designed to obtain predictions of performance in stressful combat situations may not lead to responses which correlate with day-to-day peacetime performance. Question developers and respondents should have a common understanding of the situations for which behavior is being predicted, and criterion measures should be obtained in the same situation. Questions should also be specific as to the action being predicted and the target of that action. Research has shown that attitudes correspond more closely to behavior as actions and targets are specified in greater detail (Ajzen and Fishbein, 1977). The implication of this finding for subjective appraisals of proficiency is that the action or behavior to be predicted should be specified in detail, along with a clear definition of when the action is completed and what the result is.

Relevant to this discussion of how to design questions to maximize the accuracy of subjective appraisals is a technique applied by Harris, Osborn, and Boldovici (1978). As described earlier, these authors found that rater agreement was typically low in studies of subjective criticality estimations. To get around this problem, they used a paired-comparison technique in which raters compared tasks to one another rather than rating each task on a numerical scale. That is, two tasks were described in a well-defined situation and subjects were asked to identify the more critical one. In this way, relative rather than absolute criticality ratings were obtained, the judgment process was simplified, and an operational definition of criticality was provided. Results showed that use of this method increased interrater reliability considerably, to higher than the .90 level in some cases. The effects of using this technique on the accuracy or predictive validity of criticality estimations was not directly addressed, but an approach which increases the reliability of subjective appraisals would be expected to also have a positive impact upon validity. One operational problem with this approach is the extent to which complete pairings of tasks can be presented for comparison. With more than a few (six or eight) tasks being evaluated, the number of pairs becomes so large as to preclude presentation of them all to all raters. In this case, some method of partial pairing must be used, and the best way to do this is not always clear. So this technique would best be utilized when a small number of tasks are being compared. It could easily be adapted to situations where the performance proficiency, frequency, or difficulty of specific tasks is being appraised, as well as the criticality.

Raters' Experiences

Another major variable impacting upon the accuracy of subjective appraisals is the extent to which raters share common experiences. This variable has most commonly been addressed in terms of training provided to raters before they provide subjective appraisals. Cascio (1978) reviewed the effects of such training programs and concluded that training for raters is most beneficial when it includes practice with the specific rating scales to be used, discussions of errors commonly made by raters, and emphasis upon distinguishing among the different aspects or dimensions of a situation. Research results

indicate that training programs designed in accord with these recommendations reduce the amount of halo effect and other errors in subjective ratings. Bergman and Siegel (1972) concluded that training programs are effective to the extent that they eliminate idiosyncrasies in the way raters observe their own or others' behavior. There are also indications that the degree or type of training impacts upon its effect. For example, Bernardin and Walter (1977) found that one hour of training on the nature of psychometric errors resulted in significantly less halo error in subsequently obtained ratings. But exposure to the scale to be utilized in addition to one hour of training resulted in less leniency error and higher interrater reliability, in addition to reduced halo error. So training in making subjective appraisals can be expected to have a positive impact upon their accuracy. This training should include a general discussion of the types of errors commonly made and experience with the specific rating scale to be used. If a large number of subjective appraisals are being collected over a long period of time, training should be provided during the rating period as well as before it. Research summarized by Frick and Semmel (1978) has shown that reliability of ratings may decrease as a function of time since training.

Other Characteristics of Raters

Shrauger and Osberg (1981) have summarized several other characteristics of raters which may influence the accuracy of subjective appraisals. One important consideration is whether raters have the intellectual or cognitive capacity to effectively appraise their own and others' performance. Most studies of the accuracy of subjective appraisals have used subjects of above average educational and intellectual levels. These studies have generally found low accuracy, and the accuracy might be even less for samples of soldiers, many of whom have not completed a high school education. This hypothesis is supported by Gorsuch, Henighan, and Barnard (1972), who found that the reliability of a scale depended upon the reading ability of the raters. Errors of measurement were found to be small for good readers, but were large for poor readers. Further research is needed to address the relationship between level of education and ability to make accurate subjective appraisals.

Another individual characteristic which has been found to influence the accuracy of self-appraisals is the degree of raters' self-consciousness or self-awareness. While this variable may be difficult to operationalize, it could perhaps be delineated in terms of experiences on specific tasks. Individuals would be expected to provide more accurate subjective appraisals for tasks with which they have extensive experience, and they should never be asked to appraise tasks with which they have little or no experience. Data supporting this point have been reported by Primoff (1979). He found that job applicants were moderately accurate in self-appraising their abilities on familiar tasks, such as spelling, but were not accurate on less familiar tasks, such as comparing names and numbers. Ash (1980) reported similar results for typing tasks. Supervisory appraisals should also be expected to be more accurate for familiar tasks on which performance has been observed frequently, as shown by the research of MacLane (1977) described earlier. The consistency of the appraised individuals' behavior will also impact upon appraisal accuracy; such accuracy should be higher with tasks for which behavior is consistent rather than highly variable. Consistent experience with tasks will not facilitate

appraisals unless raters can remember it. Recall of relevant previous experience should be facilitated before appraisals are given. This can be done by asking raters to review their behavior in previous relevant situations or by providing them with memory cues, such as descriptions of the tasks being addressed and situations in which they are commonly performed.

Motivation is another factor which can influence the accuracy of subjective appraisals. The need for accuracy should be strongly emphasized in instructions provided before ratings are collected. The accuracy of at least a selected sample of subjective ratings should be checked against objective criteria, such as performance test results. Raters should be informed that such a check will be conducted, in order to maximize their desire for accuracy.

In summary, while the degree of accuracy of subjective appraisals is yet unknown, it can be maximized through the application of methodologically sound data collection approaches. Some of these techniques were described above and will be summarized as recommendations in the next section. Further research is needed to determine the exact relationship of these approaches to the accuracy of subjective appraisals. Using these techniques to collect subjective appraisals in conjunction with the collection of more objective comparative data will provide many of the data that are needed.

CONCLUSIONS/RECOMMENDATIONS

The data reviewed in this paper lead to at least three major conclusions with respect to the accuracy of subjective appraisals. The first of these is that adequate data are not yet available to determine either the absolute accuracy of subjective appraisals or the relative accuracy of different appraisal sources. The biggest problem here is the general lack of objective criteria to which subjective data can be compared. In many studies, subjective ratings have been compared to other ratings or to data which only approximate objective criterion data, such as written test results. When ratings from different sources have been compared to each other, results show that self-appraisals differ somewhat from peer and supervisory appraisals. But ratings have not in general been compared to sufficiently objective criteria to allow definitive statements on their accuracy or predictive validity. Research is badly needed which allows comparison of subjective ratings or predictions to relatively objective sets of criterion data, such as results of "hands-on" performance tests.

The second major conclusion is that the limited research which has directly addressed the accuracy of subjective appraisals has in general not found it to be high. Results for appraisals of the performance proficiency, frequency, difficulty, and criticality of specific tasks all support this conclusion. Various types of psychometric errors have commonly been found in subjective appraisals. The general lack of interrater reliability limits the amount of accuracy or validity that can be expected in subjective appraisals. People have difficulty distinguishing among the various aspects or dimensions of an appraisal situation, which often leads to halo effects. A leniency error or positive bias has frequently been found, especially in self-appraisals. Before conclusions are drawn based upon subjective appraisals in any situation, the accuracy of the data should be checked. This check should involve a comparison

of subjective data with independently gathered data that are as objective in nature as possible.

The final conclusion is that while the available data relating to the accuracy of subjective appraisals are not definitive, there are ways to increase this accuracy. Subjective appraisals will always be used because of the ease and economy with which they can be collected. Further research is needed, but available research results suggest several general ways in which the accuracy of subjective appraisals can be increased. These are summarized below, and their application to the collection and use of subjective appraisals is strongly recommended.

1. Integrate mutually supportive subjective appraisal methods within a feedback system. Since no appraisal method is complete and sufficient in and of itself, methods should be used to complement each other. Surveys can be used to obtain a general overview of the situation, interviews can be used to obtain more in-depth detail on specific problems, and observations and performance tests can be used as accuracy checks.

2. Ensure that question developers and subjective appraisers have a common base of understanding. These groups should share a common understanding of task elements, successful task completion, appropriate standards, and rating dimensions. If any of these factors are unclear, misleading data may result.

3. Design questions to maximize accuracy. Make the situation and behavior being addressed as explicit as possible, and specifically state the action being addressed and the target of that action. With a small number of tasks, consider using a paired-comparison rather than an absolute rating technique. With a larger number of tasks, consider asking raters to compare their own strengths and weaknesses, rather than to compare their abilities to those of others. Also, consider asking appraisers to rate their maximal rather than their typical behavior.

4. Make rating scales as explicit as possible. Phrase rating scales in terms of explicit observable measures of performance, rather than in vague, general terms such as "average," "below average," etc. Describe each rating point in terms of the behavior that it represents. Consider asking raters to provide specific examples of experiences which support their ratings.

5. Be sure the raters have had experience with the tasks rated. Give raters the option of indicating that they have not had experience with any given task, and thus cannot provide a rating for it. Be sure that supervisors have had ample opportunity to observe task performance by the people they are rating.

6. Train raters before they provide subjective appraisals. This training should include experience with the rating scales to be used, a discussion of common types of psychometric errors (halo and leniency effects), and a discussion of the dimensions of the situation being evaluated. Provide refresher training to raters if a large number of ratings are being collected over a long period of time.

7. Facilitate raters' recall of relevant experiences. Ask raters to review their previous experiences, provide them with thorough descriptions of the tasks and situations being rated, and provide any other memory cues which aid recall.

8. Make certain that appraisers have the cognitive capacity and motivation to provide accurate ratings. Be sure that they can understand the questions asked and the use of rating scales. Explain the need for accurate rating data during instructions. If the accuracy of the subjective ratings will be checked, let the raters know this.

REFERENCE NOTES

1. West, Arthur L. Jr., MG, USA (Ret.). Personal communication, March 1971.

2. Hiller, J. H. Personal communication, December 1981.

3. Goldberg, S. L. Personal communication, December 1981.

4. Hiller, J. H. Personal communication, February 1982.

REFERENCES

Ajzen, I. & Fishbein, M. Attitude-behavior relations: A theoretical analysis and review of empirical research. Psychological Bulletin, 1977, 84, 888-918.

Ash, R. A. Self-assessments of five types of typing ability. Personnel Psychology, 1980, 33, 273-282.

Bergman, B. A. & Siegel, A. I. Training evaluation and student achievement measurement: A review of the literature (Technical Report 72-3). Lowry Air Force Base, CO: Air Force Human Resources Laboratory, January 1972.

Bernardin, H. J., Beatty, R. W., & Jensen, W. The new uniform guidelines on employee selection procedures in the context of university personnel decisions. Personnel Psychology, 1980, 33, 301-316.

Bernardin, H. J. & Walter, C. S. Effect of rater-training and diary-keeping on psychometric error in ratings. Journal of Applied Psychology, 1977, 62, 64-69.

Burnside, B. L. Field performance feedback - A problem review (Research Report 1323). Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences, August 1981. (ADA 134 388)

Carver, R. P. Implications of a new technique for measuring the understanding gained from reading for non-residential programs. Washington, D. C.: American Institutes for Research, 1972. (ERIC Document Reproduction Service No. ED 064 383 .)

Cascio, W. F. Applied psychology in personnel management. Reston, VA: Reston Publishing Company, Inc., 1978.

Cohen, P. A. Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. Review of Educational Research, 1981, 51, 281-309.

DeNisi, A. S. & Shaw, J. B. Investigation of the uses of self-reports of abilities. Journal of Applied Psychology, 1977, 62, 641-644.

Downey, R. G. & Duffy, P. J. Review of peer evaluation research (Technical Paper 342). Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences, October 1978. (ADA 061 780)

Dyer, F. N. & Hilligoss, R. E. Using an assessment center to predict field leadership performance of Army officers and NCO's (Technical Paper 372). Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences, May 1979.

Flavell, J. H. Developmental studies of mediated memory. In H. W. Reese & L. P. Lipsitt (Eds.), Advances in child development and behavior (Vol. 5). New York: Academic Press, 1970.

Frick, T. & Semmel, M. I. Observer agreement and reliabilities of classroom observational measures. Review of Educational Research, 1978, 48, 157-184.

Gardiner, J. M. & Klee, H. Memory for remembered events: An assessment of output monitoring in free recall. Journal of Verbal Learning and Verbal Behavior, 1976, 15, 227-233.

Gilbert, A. C. F. & Downey, R. G. Validity of peer ratings obtained during Ranger training (Technical Paper 344). Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences, October 1978. (ADA 061 576)

Gorsuch, R. L., Henighan, R. P., & Barnard, C. Locus of control: An example of dangers in using children's scales with children. Child Development, 1972, 43, 579-590.

Hall, E. R., Denton, C. F., & Zajkowski, M. M. A comparative assessment of three methods of collecting training feedback information (TAEG Report No. 64). Orlando, FL: Training Analysis and Evaluation Group, December 1978.

Harris, J. H., Osborn, W. C., & Boldovici, J. A. A paired-comparison approach for estimating task criticality. In Osborn, W. C., Ford, J. P., Campbell, C. H., Campbell, R. C., Harris, J. H., & Boldovici, J. A. Military testing: Knowledge and skills (Professional Paper 4-78). Alexandria, VA: Human Resources Research Organization, February 1978.

Harris, J. H., Campbell, C. H., & Osborn, W. C. An attempt to identify indicators of competence on mechanical maintenance tasks (Final Report 79-1). Alexandria, VA: Human Resources Research Organization, January 1979.

Hermont, I. What is the Army getting for its training dollar? Army, 1977 (June), 34-38.

Hiller, J. H. Learning from prose text: Effects of readability level, inserted question difficulty, and individual differences. Journal of Educational Psychology, 1974, 66, 202-211.

Hiller, J. H. A methodology for estimating the cost-effectiveness of alternative pretesting procedures (Technical Report 502). Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences, November 1980. (ADA 115 877)

Holzbach, R. L. Rater bias in performance ratings: Superior, self-, and peer ratings. Journal of Applied Psychology, 1978, 63, 579-588.

Hook, C. M. & Rosenshine, B. V. Accuracy of teacher reports of their classroom behavior. Review of Educational Research, 1979, 49, 1-12.

Johnson, C. A., Tokunaga, H. T., & Hiller, J. H. Validation of a job analysis questionnaire against intensive observation. Paper presented at the Military Testing Association Conference, Toronto, October 1980.

Kahneman, D. & Tversky, A. On the psychology of prediction. Psychological Review, 1973, 80, 237-251.

Kane, J. S. & Lawler, E. E. Methods of peer assessment. Psychological Bulletin, 1978, 85, 555-586.

Kroll, N. E. A. & Kellicutt, M. H. Short-term recall as a function of covert rehearsal and of intervening task. Journal of Verbal Learning and Verbal Behavior, 1972, 11, 196-204.

Lachman, J. L., Lachman, R., & Thronesbery, C. Metamemory through the adult life span. Developmental Psychology, 1979, 15, 543-551.

Lammlein, S. E. & Borman, W. C. Peer rating research: Annotated bibliography (Technical Report 79-9). Brooks Air Force Base, TX: Air Force Human Resources Laboratory, June 1979.

Lieberman, D. A. Behaviorism and the mind: A (limited) call for a return to introspection. American Psychologist, 1979, 34, 319-333.

Levine, E. L. Introductory remarks for the symposium "Organizational applications of self-appraisal and self-assessment: Another look." Personnel Psychology, 1980, 33, 259-262.

Levine, E. L., Flory, A., & Ash, R. A. Self-assessment in personnel selection. Journal of Applied Psychology, 1977, 62, 428-435.

MacLane, C. N. Promotion evaluation for inter-organizational referral: A behavioral expectation approach. Paper presented at the Military Testing Association Conference, San Antonio, October 1977.

Medlin, S. M. & Thompson, P. Evaluator rating of unit performance in field exercises: A multidimensional scaling analysis (Technical Report 438). Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences, April 1980. (ADA 089 264)

Meyer, H. H. Self-appraisal of job performance. Personnel Psychology, 1980, 33, 291-296.

Mitchell, S. K. Interobserver agreement, reliability, and generalizability of data collected in observational studies. Psychological Bulletin, 1979, 86, 376-390.

Moreland, R., Miller, J., & Laucka, F. Academic achievement and self-evaluation of academic performance. Journal of Educational Psychology, 1981, 73, 335-344.

Nisbett, R. E. & Wilson, T. D. Telling more than we can know: Verbal reports on mental processes. Psychological Review, 1977, 84, 231-259.

Pourchot, L. & Lanning, F. The self-concept as predictor of scores on the Pourchot Mechanical Manipulation Test. Journal of the Association for the Study of Perception, 1979, 14, 6-11.

Primoff, E. S. The use of self-assessments in examining (Professional Series 79-1). Washington, DC: Office of Personnel Management, Personnel Research and Development Center, April 1979.

Primoff, E. S. The use of self-assessments in examining. Personnel Psychology, 1980, 33, 283-290.

Robinson, J. A. & Kulp, R. A. Knowledge of prior recall. Journal of Verbal Learning and Verbal Behavior, 1970, 9, 84-86.

Rose, A. M. & Wheaton, G. R. Performance effectiveness in combat job specialties (Final Report 5173-66200). Washington, D. C.: American Institutes for Research, May 1978.

Ryan-Jones, D. L. A comparison of expert ratings of task difficulty with an independent criterion (Technical Report 418). Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences, November 1979. (ADA 082 016)

Schendel, J. D. & Hagman, J. D. On sustaining procedural skills over a prolonged retention interval. Journal of Applied Psychology, in press.

Shaughnessy, J. J. Confidence-judgment accuracy as a predictor of test performance. Journal of Research in Personality, 1979, 13, 505-514.

Shavelson, R. & Dempsey-Atwood, N. Generalizability of measures of teaching behavior. Review of Educational Research, 1976, 46, 553-611.

Shields, J. L., Goldberg, S. L., & Dressel, J. D. Retention of basic soldiering skills (Research Report 1225). Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences, September 1979. (ADA 075 412)

Shrauger, J. S. & Osberg, T. M. The relative accuracy of self-predictions and judgments by others in psychological assessment. Psychological Bulletin, 1981, 90, 322-351.

Shvern, U. Field evaluation of the combat commander's guide to aerial surveillance and reconnaissance resources (Technical Paper 380). Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences, July 1979. (ADA 075 422)

Smith, E. R. & Miller, F. D. Limits on perception of cognitive processes: A reply to Nisbett and Wilson. Psychological Review, 1978, 85, 355-362.

Thornton, G. C. Psychometric properties of self-appraisals of job performance. Personnel Psychology, 1980, 33, 263-272.

Tulving, E. Episodic and semantic memory. In E. Tulving and W. Donaldson (Eds.), Organization of memory. New York: Academic Press, 1972.

Turney, J. R. & Cohen, S. L. Perceived work effort as time devoted to an activity (Technical Paper 337). Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences, September 1978. (ADA 062 411)

US Army Field Manual 101-5, Staff officers field manual: Staff organization and procedure. Washington, D. C.: Headquarters, Department of the Army, July 1972.

US Army Regulation 220-1, Unit status reporting. Washington, D. C.: Headquarters, Department of the Army, June 1978.

US Army Training and Doctrine Command Pamphlet 350-30, Interservice procedures for instructional systems development. Fort Benning, GA: Combat Arms Training Board, August 1975.

US Army Training and Doctrine Command Draft Regulation 350-7, A systems approach to training. Fort Monroe, VA: Headquarters, US Army Training and Doctrine Command, undated.

US Army Training Study. Fort Belvoir, VA, 1978.

van Rijn, P. Self-assessment for personnel examining: An overview (Personnel Research Report 80-14). Washington, D. C.: Office of Personnel Management, Personnel Research and Development Center, June 1980.

van Rijn, P. Self-assessment in personnel selection and placement. Paper presented at the Military Testing Association Conference, Arlington, VA, October 1981.

Wills, T. A. Perceptions of clients by professional helpers. Psychological Bulletin, 1978, 33, 344-358.